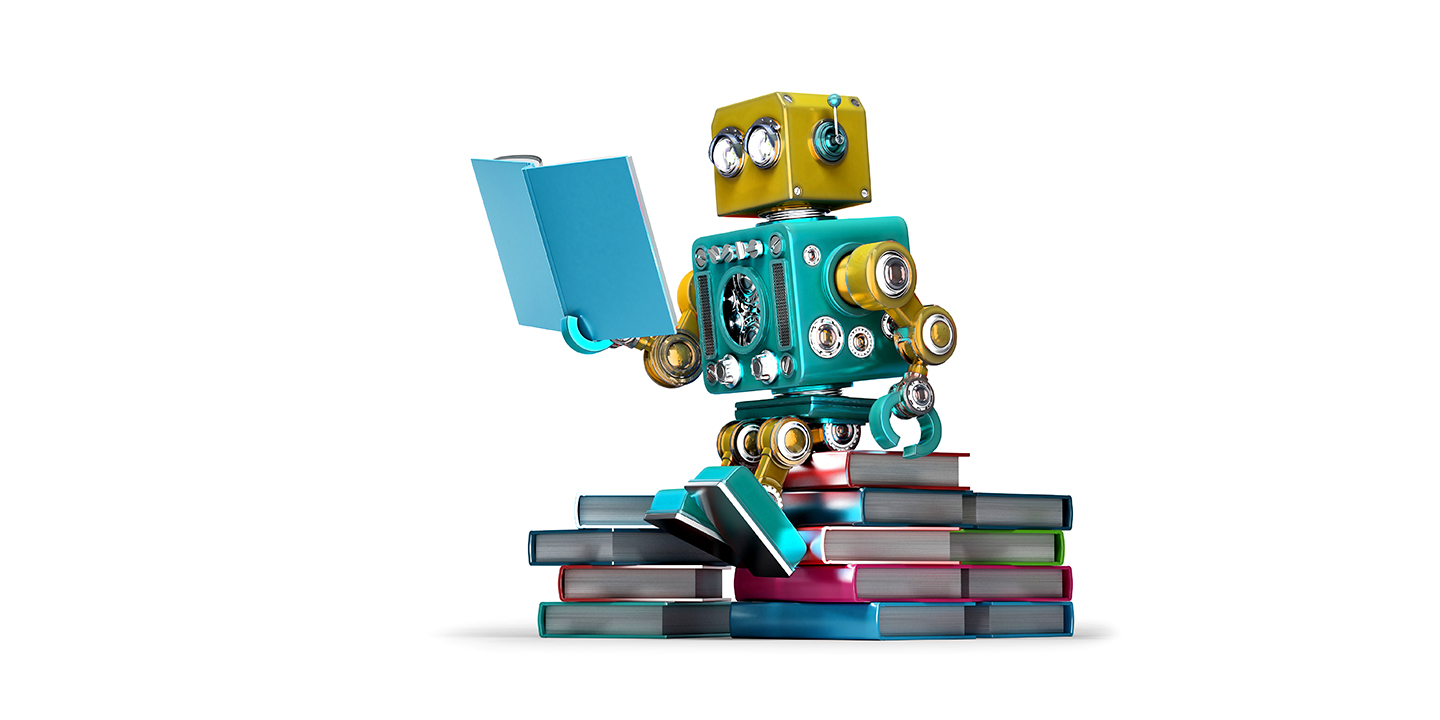
**Μηχανική Μάθηση Εργασία 1**



**Μπόζας Αριστείδης**

**ΑΜ:740**

**Γιάννης Δερμετζόγου**

**ΑΜ:**

**Περιεχόμενα**

[1. Μέρος Α 4](#_Toc511454016)

[2. Μέρος Β 4](#_Toc511454017)

[3. Μέρος Γ 5](#_Toc511454018)

[Παράρτημα κώδικα 1 5](#_Toc511454019)

[Παράρτημα κώδικα 2 5](#_Toc511454020)

[Παράρτημα κώδικα 3 5](#_Toc511454021)

Κατάλογων Πινάκων

Κατάλογων Εικόνων

# Μέρος Α

Στο πρώτο μέρος της εργασίας εξετάστηκε το αντικείμενο των τεχνικών πολλαπλών μοντέλων πρόβλεψης σε συνδυασμό με το αντικείμενο της συγκριτικής αξιολόγησης μεταξύ αλγορίθμων.

Τα 10 datasets που χρησιμοποιήθηκαν από το [UCI](http://archive.ics.uci.edu/ml/index.php) repository είναι τα εξής:

* [Iris](https://archive.ics.uci.edu/ml/datasets/iris)
* [Wine](https://archive.ics.uci.edu/ml/datasets/wine)
* [Wdbc](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))
* [Balance-scale](http://archive.ics.uci.edu/ml/datasets/balance+scale)
* [Hayes roth](https://archive.ics.uci.edu/ml/datasets/Hayes-Roth)
* [Haberman survival](https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival)
* [Liver disorder](https://archive.ics.uci.edu/ml/datasets/liver+disorders)
* [Data bank authedication](https://archive.ics.uci.edu/ml/datasets/banknote+authentication)
* [Ionosphere](https://archive.ics.uci.edu/ml/datasets/ionosphere)
* [Cmc](https://archive.ics.uci.edu/ml/machine-learning-databases/cmc/)

Οι τεχνικές που χρησιμοποιήθηκαν για τα πολλαπλά μοντέλα πρόβλεψη είναι οι εξής:

* **Manipulating the training examples:** Oι τεχνικές του bagging και boosting.
* **Manipulating the target variable:** Oι τεχνικές του OnevsOne και OnevsRest.
* **Injecting randomness :** Tο ensemble μοντέλο RandomForest.
* **Manipulating Features :** Τυχαία επιλογή του των Feature και των παραδειγμάτων εκπαίδευσης με την τεχνική RandomPatches.

Όσο αναφορά το αντικείμενο της συγκριτικής αξιολόγησης μεταξύ των αλγορίθμων που εκτελέστηκαν φαίνονται παρακάτω:

# Μέρος Β

Το δεύτερο μέρος της εργασία εξετάστηκε το πρόβλημα του διαφορετικού κόστους στο σύνολο δεδομένου [heart](http://archive.ics.uci.edu/ml/datasets/statlog+(heart)) .H βιβλιοθήκη που χρησιμοποιήθηκε ήταν η **Costcla,sklearn** της python.

To cost matrix αυτού του συνόλου δεδομένου είναι το εξής:

|  |  |  |
| --- | --- | --- |
|  | **Actual absence** | **Actual presence** |
| **Absence** | 0 | 1 |
| **Presence** | 5 | 0 |

Πίνακας : Cost matrix του συνόλου δεδομένου heart

Οι μέθοδοι που χρησιμοποιήθηκαν και αναφέρονται παρακάτω συνδιάστηκαν με τους αλγορίθμους μάθησης Random Forest, Linear SVM,Naive Bayes:

* Δίνοντας έμφαση στα παραδείγματα με το μεγαλύτερο κόστος(**CostSampling** [Oversampling, RejectionSampling) )
* Ελαχιστοποίηση αναμενόμενου κόστους εκτιμήσεων (**ThresholdOptimization,BayesMinimumRiskClassifier)**
* Τροποποιημένη cost sensitive ταξινομητές( **CostSensitiveRandoForestClassifier**)

Επειδή οι πιθανότητες παίζουν μεγάλο ρόλο στο πρόβλημα του διαφορετικού κόστους έγινε ένα calibration στις πιθανότητες των αλγορίθμων μάθησης Random Forest, Linear SVM,Naive Bayes με τη χρήση της βιβλιοθήκης **sklearn.calibration.**

|  |  |
| --- | --- |
| **Αλγόριθμος** | **Costloss** |
| **RF** | 59.0 |
| **RF - O** | 41.0 |
| **RF - R** | 41.0 |
| **RF - U** | 53.0 |
| **LSVM** | 60.0 |
| **LSVM - O** | 42.0 |
| **LSVM - R** | 37.0 |
| **LSVM - U** | 61.0 |
| **GNB** | 54.0 |
| **GNB - O** | 45.0 |
| **GNB – R** | 32.0 |
| **GNB - O** | 50.0 |
| **RF - BMR** | 28.0 |
| **RF - TO** | 165.0 |
| **LSVM - BMR** | 33.0 |
| **LSVM - TO** | 37.0 |
| **GNB - BMR** | 33.0 |
| **GNB - TO** | 33.0 |
| **RFC** | 43 |

# Μέρος Γ

# Παράρτημα κώδικα 1

# Παράρτημα κώδικα 2

# Παράρτημα κώδικα 3